Rule Induction with Grouping Target Concepts based on Rough Sets

Shusaku Tsumoto [1,2]

Department of Medical Informatics,
Shimane Medical University, School of Medicine,
Enya-cho Izumo City, Shimane 693-8501 Japan

Abstract

One of the most important problems with rule induction methods is that they cannot extract rules, which plausibly represent experts' decision processes. In this paper, the characteristics of experts' rules are closely examined and a new approach to extract plausible rules is introduced, which consists of the following three procedures. First, the characterization of decision attributes (given classes) is extracted from databases and the concept hierarchy for given classes is calculated. Second, based on the hierarchy, rules for each hierarchical level are induced from data. Then, for each given class, rules for all the hierarchical levels are integrated into one rule. The proposed method was evaluated on a medical database, the experimental results of which show that induced rules correctly represent experts' decision processes.

Key words: Rule induction, grouping, coverage, Rough Sets, Granular Computing.

1 Introduction

One of the most important problems in data mining is that extracted rules are not easy for domain experts to interpret. One of its reasons is that conventional rule induction methods[5] cannot extract rules, which plausibly represent experts' decision processes[7]: the description length of induced rules is too short, compared with the rules acquired from domain experts. For example, rule induction methods, including C4.5[4] and PRIMEROSE[7], induce the following common rule for muscle contraction headache from databases on differential diagnosis of headache:

[1] This work was supported by the Grant-in-Aid for Scientific Research (13131208) on Priority Areas (No.759) "Implementation of Active Mining in the Era of Information Flood" by the Ministry of Education, Science, Culture, Sports, Science and Technology of Japan.
[2] Email: tsumoto@computer.org

$[location = whole] \wedge$[Jolt Headache $= no$] $\wedge$[Tenderness of M1 $= yes$]

$\rightarrow$ muscle contraction headache.

This rule is shorter than the following rule given by medical experts.

[Jolt Headache $= no$]

$\wedge$([Tenderness of M0 $= yes$] $\vee$[Tenderness of M1 $= yes$]

$\vee$[Tenderness of M2 $= yes$])

$\wedge$[Tenderness of B1 $= no$] $\wedge$[Tenderness of B2 $= no$]

$\wedge$[Tenderness of B3 $= no$]

$\wedge$[Tenderness of C1 $= no$] $\wedge$[Tenderness of C2 $= no$]

$\wedge$[Tenderness of C3 $= no$] $\wedge$[Tenderness of C4 $= no$]

$\rightarrow$ muscle contraction headache

where many attribute-value pairs are added.

These results suggest that conventional rule induction methods do not reflect a mechanism of knowledge acquisition of medical experts.

In this paper, the characteristics of experts' rules are closely examined and a new approach to extract plausible rules is introduced, which consists of the following three procedures. First, the characterization of decision attributes (given classes) is extracted from databases and the concept hierarchy for given classes is calculated. Second, based on the hierarchy, rules for each hierarchical level are induced from data. Finally, for each given class, rules for all the hierarchical levels are integrated into one rule.

The paper is organized as follows. Section 2 discusses the background of this study. Section 3 and 4 introduces rough sets and a characterization set. Section 5 gives an algorithm for rule induction. Section 6 shows an illustrative example. Finally, Section 7 concludes this paper.

2 Background: Problems with Rule Induction

As shown in the introduction, rules acquired from medical experts are much longer than those induced from databases the decision attributes of which are given by the same experts. This is because rule induction methods generally search for shorter rules. One of the main reasons why rules are short is that these patterns are generated only by a single criterion, such as high accuracy or high information gain. The comparative studies[7,8] suggest that experts should acquire rules not only by a single criterion but by several different diagnostic criteria. Those characteristics of medical experts' rules can be fully examined not by comparing between those rules for the same class, but by comparing experts' rules with those for another class[7]. For example, the classification rule for muscle contraction headache given in Section 1 is very similar to the following classification rule for diseaseof cervical spine:

[Jolt Headache = no]
$\wedge$([T enderness of M0 =yes] $\vee$[Tenderness of M1 = yes]
$\vee$[T enderness of M2 =yes])
$\wedge$([T enderness of B1 =yes] $\vee$[Tenderness of B2 = yes]
$\vee$[Tenderness of B3 = yes]
$\vee$[Tenderness of C1 = yes] $\vee$[Tenderness of C2 = yes]
$\vee$[Tenderness of C3 = yes] $\vee$[Tenderness of C4 = yes])
$\rightarrow$ disease of cervical spine

The differences between these two rules are attribute-value pairs, from tenderness of B1 to C4. Thus, these two rules are composed of the following three blocks:

$A_1 \wedge A_2 \wedge \neg A_3 \rightarrow muscle\ contraction\ headache$
$A_1 \wedge A_2 \wedge A_3 \rightarrow disease\ of\ cervical\ spine$,

where A_1, A_2 and A_3 are given as the following formulae:
A_1 = [Jolt Headache = no], A_2 = [Tenderness of M0 = yes] $\vee$ [T endernessof $M1 = yes$] $\vee$ [Tenderness of M2 = yes], and A_3 = [Tenderness of C1 = no] $\wedge$ [Tenderness of C2 = no] $\wedge$ [T endernessof C3 = no] $\wedge$ [T endernessof C4 = no]. The first two blocks (A_1 and A_2) and the third one (A_3) represent the different types of differential diagnosis. The first one A_1 shows the discrimination between muscular type and vascular type of headache. Then, the second part shows that between headache caused by neck and head muscles. Finally, the third formula A_3 is used to make a differential diagnosis between muscle contraction headache and disease of cervical spine. Thus, medical experts first select several diagnostic candidates, which are very similar to each other, from many diseases and then make a final diagnosis from those candidates.

3 Rough Set Theory and Probabilistic Rules

3.1 Rough Set Notations

In the following sections, the following notations introduced by Grzymala-Busse and Skowron[6], are used which are based on rough set theory[2]. These notations are illustrated by a small database shown in Table 1, collecting the patients who complained of headache.

Let U denote a nonempty, finite set called the universe and A denote a nonempty, finite set of attributes, i.e., $a : U \rightarrow V_a$ for $a \in A$, where V_a is called the domain of a, respectively. Then, a decision table is defined as an information system, $A = (U, A \cup \{d\})$. For example, Table 1 is an information system with $U = \{1, 2, 3, 4, 5, 6\}$ and $A = \{age, location, nature, prodrome, nausea, M1\}$ and $d = class$. For $location \in A$, $V_{location}$ is defined as $\{occular, lateral, whole\}$.

The atomic formulae over $B \subseteq A \cup \{d\}$ and V are expressions of the form

Table 1
A small example of a database

No.	loc	nat	his	prod	jolt	nau	M1	M2	class
1	occular	per	per	0	0	0	1	1	m.c.h.
2	whole	per	per	0	0	0	1	1	m.c.h.
3	lateral	thr	par	0	1	1	0	0	common.
4	lateral	thr	par	1	1	1	0	0	classic.
5	occular	per	per	0	0	0	1	1	psycho.
6	occular	per	subacute	0	1	1	0	0	i.m.l.
7	occular	per	acute	0	1	1	0	0	psycho.
8	whole	per	chronic	0	0	0	0	0	i.m.l.
9	lateral	thr	per	0	1	1	0	0	common.
10	whole	per	per	0	0	0	1	1	m.c.h.

Definition. loc: location, nat: nature, his:history,
Definition. prod: prodrome, nau: nausea, jolt: Jolt headache,
M1, M2: tenderness of M1 and M2, 1: Yes, 0: No, per: persistent,
thr: throbbing, par: paroxysmal, m.c.h.: muscle contraction headache,
psycho.: psychogenic pain, i.m.l.: intracranial mass lesion, common.:
common migraine, and classic.: classical migraine.

$[a = v]$, called descriptors over B, where $a \in B$ and $v \in V_a$. The set $F(B, V)$ of formulas over B is the least set containing all atomic formulas over B and closed with respect to disjunction, conjunction and negation. For example, $[location = occular]$ is a descriptor of B.

For each $f \in F(B, V)$, f_A denote the meaning of f in A, i.e., the set of all objects in U with property f, defined inductively as follows.

(i) If f is of the form $[a = v]$ then, $f_A = \{s \in U | a(s) = v\}$

(ii) $(f \wedge g)_A = f_A \cap g_A$; $(f \vee g)_A = f_A \vee g_A$; $(\neg f)_A = U - f_a$

For example,$f = [location = occular]$ and $f_A = \{1, 5, 6, 7\}$. As an example of a conjunctive formula, $g = [location = occular] \wedge [nausea = no]$ is a descriptor of U and g_A is equal to $\{1, 5\}$.

By the use of the framework above, classification accuracy and coverage, or true positive rate is defined as follows.

Definition 3.1 Let R and D denote a formula in $F(B, V)$ and a set of objects whose decision class is d. Classification accuracy and coverage(truepositive rate) for $R \rightarrow d$ is defined as:

$$\alpha_R(D) = \frac{|R_A \cap D|}{|R_A|} (= P(D|R)), \text{ and } \kappa_R(D) = \frac{|R_A \cap D|}{|D|} (= P(R|D)),$$

where $|S|$, $\alpha_R(D)$, $\kappa_R(D)$ and P(S) denote the cardinality of a set S, a classification accuracy of R as to classification of D and coverage(a true positive rate of R to D), and probability of S, respectively.

In the above example, when R and D are set to $[nau = 1]$ and $[class = common]$, $\alpha_R(D) = 2/5 = 0.4$ and $\kappa_R(D) = 2/2 = 1.0$.

Finally, we define the partial order of formulae as follows:

Definition 3.2 Let R_i and R_j be the formulae in $F(B, V)$ and let $A(R_i)$ denote a set whose elements are the attribute-value pairs of the form $[a, v]$

included in R_i. If $A(R_i) \subseteq A(R_j)$, then we represent this relation as:

$$R_i \preceq R_j.$$

3.2 Probabilistic Rules

According to the definitions, probabilistic rules with high accuracy and coverage are defined as:

$$R \stackrel{\alpha,\kappa}{\rightarrow} d \ s.t. \ R = \vee_i R_i = \vee \wedge_j [a_j = v_k], \ \alpha_{R_i}(D) \geq \delta_\alpha \text{ and } \kappa_{R_i}(D) \geq \delta_\kappa,$$

where δ_α and δ_κ denote given thresholds for accuracy and coverage.

4 Characterization Sets

4.1 Characterization Sets

In order to model medical reasoning, a statistical measure, coverage plays an important role in modeling, which is a conditional probability of a condition (R) under the decision D($P(R|D)$). Let us define a characterization set of D, denoted by $L(D)$ as a set, each element of which is an elementary attribute-value pair R with coverage being larger than a given threshold, δ_κ. That is,

Definition 4.1 Let R denote a formula in $F(B, V)$. Characterization sets of a target concept (D) is defined as:

$$L_{\delta_\kappa}(D) = \{R|\kappa_R(D) \geq \delta_\kappa\}.$$

Then, three types of relations between characterization sets can be defined as follows:

Independent type:	$L_{\delta_\kappa}(D_i) \cap L_{\delta_\kappa}(D_j) = \phi$,
Boundary type:	$L_{\delta_\kappa}(D_i) \cap L_{\delta_\kappa}(D_j) \neq \phi$, and
Positive type:	$L_{\delta_\kappa}(D_i) \subseteq L_{\delta_\kappa}(D_j)$.

All three definitions correspond to the negative region, boundary region, and positive region, respectively, if a set of the whole elementary attribute-value pairs will be taken as the universe of discourse.

4.2 Characteristics

We consider the special case of characterization sets in which the thresholds of coverage is equal to 1.0. That is,

$$L_{1.0}(D) = \{R_i|\kappa_{R_i}(D) = 1.0\}$$

For example, $[location = nat] \vee [location = whole]$, $[nat = per]$ and $[his = per]$ are elements of $L_{1.0}(m.c.h.)$. This characterization set has several interesting characteristics.

Theorem 4.2 *Let R_i and R_j two formulae in $L_{1.0}(D)$ such that $R_i \preceq R_j$. Then,*

$$\alpha_{R_i} \leq \alpha_{R_j}.$$

Thus, when we collect the formulae whose values of coverage are equal to 1.0, the sequence of conjunctive formulae corresponds to the sequence of increasing chain of accuracies.

For example, $[nat = per]$ and $[his = per]$ are elements of $L_{1.0}(m.c.h.)$ and those accuracies are: 3/7 and 3/5. Then, since the meaning of $([nat = per] \wedge [his = per]$ is equal to $[1, 2, 5, 10]$, the accuracy of $[nat = per] \wedge [his = per]$ is 3/4.

Since $\kappa_R(D) = 1.0$ means that the meaning of R covers all the samples of D, its complement $U - R_A$, that is, the meaning of $\neg R$ does not cover any samples of D. Especially, when R consists of the formulae with the same attributes, it can be viewed as the generation of the coarsest partitions. Thus,

Theorem 4.3 *Let R be a formula in $L_{1.0}(D)$ such that $R = \vee_j [a_i = v_j]$. Then, R and $\neg R$ give the coarsest partition for a_i in which the meaning of R includes D.* □

From the propositions 4.2 and 4.3, the next theorem holds.

Theorem 4.4 *Let A consist of $\{a_1, a_2, \cdots, a_n\}$ and R_i be a formula in $L_{1.0}(D)$ such that $R_i = \vee_j [a_i = v_j]$. Then, a sequence of conjunctive formulae $F(k) = \wedge_{i=1}^{k} R_i$ $(k \leq n)$ gives a sequence which increases the accuracy.* □

5 Rule Induction with Grouping

As discussed in Section 2, When the coverage of R for a target concept D is equal to 1.0, R is a necessity condition of D. That is, a proposition $D \rightarrow R$ holds and its contrapositive $\neg R \rightarrow \neg D$ holds. Thus, if R is not observed, D cannot be a candidate of a target concept. If two target concepts have a common formula R whose coverage is equal to 1.0, then $\neg R$ supports the negation of two concepts, which suggests that these two concepts can be grouped into the generalized concept. Furthermore, if two target concepts have similar formulae $R_i, R_j \in L_{1.0}(D)$, they are very close to each other with respect to the negation of two concepts. In this case, the attribute-value pairs in the intersection of $L_{1.0}(D_i)$ and $L_{1.0}(D_j)$ give a characterization set of the concept that unifies D_i and D_j, denoted by D_k. Then, compared with D_k and other target concepts, classification rules for D_k can be obtained. When we have a sequence of grouping, classification rules for a given target concepts are de-

```
procedure Total Process;
  var inputs
    L_D : List; /* A list of Target Concepts */
  begin
    Calculate a set of characterization set L_c;
      Calculate a set of intersection L_id;
        Calculate a list of similarity measures L_s;
          Calculate a list of grouping L_g; (Fig. 2)
        Induce a set of rules for L_g: L_r; (Fig. 3)
      Combine Rules in L_r for eac hD_i;
  end {Total Process}
```

Fig. 1. An Algorithm for Total Process

fined as a sequence of subrules each of which shows the discrimination of a generalized concept.

From these ideas, a rule induction algorithm with grouping target concepts can be described as Figure 1. This algorithm first calculates $L_{1.0}(D_i)$ for $\{D_1, D_2, \cdots, D_k\}$. Second, from the list of characterization sets, it calculates the intersection between $L_{1.0}(D_i)$ and $L_{1.0}(D_j)$ and stores it in to L_{id}. Third, the procedure calculates the similarity (matching number)of the intersections and sorts L_{id} with respect of the similarities. Finally, the algorithm chooses one intersection $(D_i \cap D_j)$ with maximum similarity (highest matching number) and group D_i and D_j in to a concept DD_i. These procedures will be continued until all the given concepts are grouped into the only one group.

```
procedure Grouping ;
  var inputs
    L_c : List; /* A list of Characterization Sets */
    L_id : List; /* A list of Intersection */ L_s : List; /* A list of Similarity */
  var outputs
    L_gr : List; /* A list of Grouping */
  var
    k : integer; L_g, L_gr : List;
  begin
    L_g := {} ; k := n; /* n: A number of Target Concepts*/
    Sort L_s with respect to similarities;
      Take a set of (D_i, D_j), L_max with maximum similarity values;
      k:= k+1;
      forall (D_i, D_j) ∈ L_max do
        begin
          Group D_i and D_j into D_k;
            L_c := L_c − {(D_i, L_1.0(D_i))}; L_c := L_c − {(D_j, L_1.0(D_j))};
              L_c := L_c + {(D_k, L_1.0(D_k))};
              Update L_id for DD_k; Update L_s;
            L_gr := ( Grouping for L_c, L_id, and L_s) ;
          L_g := L_g + {{(D_k, D_i, D_j), L_g}};
        end
    return L_g;
  end {Grouping}
```

Fig. 2. An Algorithm for Grouping

```
procedure RuleInduction ;
  var inputs
    Lc : List; /* A list of Characterization Sets */
    Lid : List; /* A list of Intersection */
    Lg : List; /* A list of grouping*/ /* {{(Dn+1,Di,Dj),{(DDn+2,.)...}}} */
               /* n: A number of Target Concepts */
  var
    Q, Lr : List;
  begin
    Q := Lg; Lr := {};
    if (Q ≠ ∅) then do
      begin
        Q := Q − first(Q); Lr := Rule Induction (Lc, Lid, Q);
      end
    (DDk, Di, Dj) := first(Q);
    if (Di ∈ Lc and Dj ∈ Lc) then do
      begin
        Induce a Rule r whic h discriminate between Di and Dj;
        r = {Ri → Di, Rj → Dj};
      end
    else do
      begin
        Search for L1.0(Di) from Lc; Search for L1.0(Dj) from Lc;
          if (i < j) then do
            begin
              r(Di) := ∨Rl∈L1.0(Dj) ¬Rl → ¬Dj; r(Dj) := ∧Rl∈L1.0(Dj) Rl → Dj;
            end
        r := {r(Di), r(Dj)};
      end
    return Lr := {r, Lr} ;
  end {Rule Induction}
```

Fig. 3. An Algorithm for Rule Induction

6 Example

Let us consider T able1 as an example for rule induction. For a similarity index, we use a matching number[1] which is defined as the cardinality of the intersection of two the sets. Also, since Table 1 has five classes, k is set to 6.

6.1 Grouping

F rom this table, the characterization set for each concept is obtained as shown in Fig 4. By using these sets, the intersection between two target concepts are calculated. Since *common* and *classic* hav e the maximum matching number, these two classes are grouped into one category, D_6. Then, the characterization of D_6 is obtained as : $D_6 = \{[loc = lateral], [nat = thr], [jolt = 1], [nau = 1], [M1 = 0], [M2 = 0]$ from Fig 5.

In the second iteration, the intersection of D_6 and others is considered as shown in Fig 6. F rom thismatrix, we hav e two possibilities of grouping: one is to group *m.c.h.* and *i.m.l.* That is, these two diseases are grouped into D_7: $D_7 = \{([loc = occular] \vee [loc = whole]), [nat = per], [prod = 0]\}$ The other one is to group D_1 and *i.m.l.*, where $D_7 = \{[jolt = 1], [M1 = 0], [M2 = 0]\}$.

In the third iteration of the former case(3_a), the in tersection is calculated as Fig 7. The following two classes, D_7 and psycho are grouped into D_8: $D_{3a} =$

{ [nat=per], [prod=0] } In the latter case(3_b), the intersection is calculated as Fig 8. The following two classes, *m.c.h.* and *psycho* are grouped into D_8: D_{8a} = { [nat=per], [prod=0] }. Fig 9 and 10 depicts the two results of grouping as two dendrograms which are mainly used for hierarchical clustering [1].

$$
\begin{aligned}
L_{1.0}(m.c.h.) = \quad & \{([loc = occular] \vee [loc = whole]), [nat = per], [his = per], \\
& [prod = 0], [jolt = 0], [nau = 0], [M1 = 1], [M2 = 1]\} \\
L_{1.0}(common) = \quad & \{[loc = lateral], [nat = thr], ([his = per] \vee [his = par]), \\
& [prod = 0], [jolt = 1], [nau = 1], [M1 = 0], [M2 = 0]\} \\
L_{1.0}(classic) = \quad & \{[loc = lateral], [nat = thr], [his = par], \\
& [prod = 1], [jolt = 1], [nau = 1], [M1 = 0], [M2 = 0]\} \\
L_{1.0}(i.m.l.) = \quad & \{([loc = occular] \vee [loc = whole]), [nat = per], \\
& ([his = subacute] \vee [his = chronic]), [prod = 0], \\
& [jolt = 1], [M1 = 1], [M2 = 1]\} \\
L_{1.0}(psycho) = \quad & \{[loc = occular], [nat = per], ([his = per] \vee [his = acute]), \\
& [prod = 0]\}
\end{aligned}
$$

Fig. 4. Characterization Sets for Table 1

6.2 Rule Induction

6.2.1 First Model for Diagnosis

Figure 9 shows one candidate of the differential diagnosis. F or the differential diagnosis of *common*. First, this model discriminates between D_6(*common* and *classic*) and D_8 (*m.c.h.*, *i.m.l.* and *psycho*). Then, *common* and *classic* within D_6 are differentiated. Thus, a classification rule for *common* is composed of two subrules: (discrimination between D_6 and D_8) and (discrimination within D_6). On the other hand, a classification rule for *m.c.h.* is composed of three subrules: (discrimination between D_6 and D_8), (discrimination between D_7 and psycho) and (discrimination within D_7).

Let us consider the first case. The first part can be obtained by the intersection in Figure 7. That is,

$$D_8 \rightarrow [nat = per] \wedge [prod = 0]$$

$$\neg[nat = per] \vee \neg[prod = 0] \rightarrow \neg D_8.$$

Then, since from Figure 4, the difference set between $L_{1.0}(common)$ and $L_{1.0}(classic)$ is $\{[prod = 1]\}$, for a classification rule for *common* within D_7 is:

$$[prod = 0] \rightarrow common.$$

Combining these two parts, the classification rule for *common* is

$$(\neg[nat = per] \vee \neg[prod = 0]) \wedge [prod = 0] \rightarrow common.$$

	m.c.h.	common	classic	i.m.l.	psycho
m.c.h.	–	{[prod=0]}	∅	{([loc=occular] ∨[loc=whole]), {[nat=per],[prod=0]}	[nat=per],[prod=0]}
common	–	–	{[loc=lateral], [nat=thr] [jolt=1], [nau=1] [M1=0],[M2=0] }	{[prod=0],[jolt=1], [M1=0], [M2=0] }	{[prod=0]}
classic	–	–	–	{[jolt=1],[M1=0], [M2=0]}	{ }
i.m.l.	–	–	–	–	{[nat=per],[prod=0]}

Fig. 5. Intersection of Two Characterization Sets (Step 2)

	m.c.h.	D_6	i.m.l.	psycho
m.c.h.	–	{}	{([loc=occular]∨[loc=whole]), {[nat=per],[prod=0]}	{[nat=per],[prod=0]}
D_6	–	–	{[jolt=1], [M1=0], [M2=0]}	{ }
i.m.l.	–	–	–	{[nat=per],[prod=0]}

Fig. 6. Intersection of Two Characterization Sets after the first Grouping (Step 3)

	D_6	D_7	psycho
D_6	–	{}	{ }
D_7	–	–	{[nat=per],[prod=0]}

Fig. 7. Intersection of Two Characterization Sets after the first Grouping (1) (Step 4a)

	m.c.h.	D_7	psycho
m.c.h.	–	{}	{[nat=per],[prod=0] }
D_7	–	{}	{ }

Fig. 8. Intersection of Two Characterization Sets after the first Grouping (2) (Step 4b)

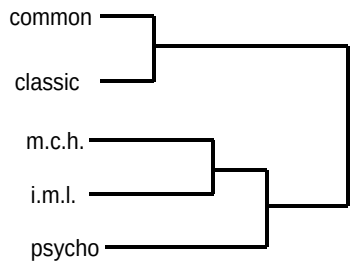

Fig. 9. Grouping by Characterization Sets (1)

After its simplification, the rule become:

$$\neg[nat = per] \rightarrow \neg common,$$

whose accuracy is equal to 2/3. In the same way, the rule for *classic* is

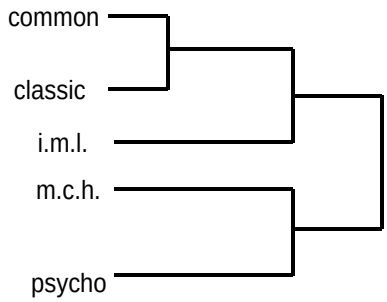

Fig. 10. Grouping by Characterization Sets (2)

obtained as:

$$\neg[nat = per] \wedge [prod = 1] \rightarrow classic.$$

6.2.2 Second Model for Diagnosis

Figure 10 shows the other candidate of the differential diagnosis. For differential diagnosis, First, this model discriminates between D_7(*common*, *classic* and *i.m.l.*) and D_8 (*m.c.h.* and *psycho*). Then, D_6 and *i.m.l.* within D_7 are differentiated. Finally, *common* and *classic* within D_7 are checked. Thus, a classification rule for *common* is composed of two subrules: (discrimination between D_7 and D_8), (discrimination between D_6 and D_7), and (discrimination within D_6).

The first part can be obtained by the intersection in Figure 7. That is,

$$D_8 \rightarrow [nat = per] \wedge [prod = 0]$$

$$\neg[nat = per] \vee \neg[prod = 0] \rightarrow \neg D_8.$$

Then, the second part can be obtained by the intersection in Figure 6. That is,

$$D_7 \rightarrow [jolt = 1] \wedge [M1 = 0] \wedge [M2 = 0]$$
$$\neg[jolt = 1] \vee \neg[M1 = 0] \rightarrow \neg D_7.$$

Finally, the third part can be obtained by the difference set between $L_{1.0}(common)$ and $L_{1.0}(classic) = \{[prod = 1]\}$.

$$[prod = 0] \rightarrow common.$$

Combining these three parts, the classification rule for *common* is

$$(\neg[nat = per] \vee \neg[prod = 0]) \wedge ([jolt = 1] \wedge [M1 = 0] \wedge [M2 = 0]) \wedge [prod = 0] \rightarrow common.$$

After its simplification, the rule is:

$$\neg[nat = per] \wedge ([jolt = 1] \wedge [M1 = 0] \wedge [M2 = 0]) \rightarrow common.$$

whose accuracy is equal to 2/3.

It is notable that the second part ($[jolt = 1] \wedge [M1 = 0] \wedge [M2 = 0]$) is redundant in this case, compared with the first model. However, from the viewpoint of characterization of a target concept, it is very important part.

7 Conclusion

In this paper, the characteristics of experts' rules are closely examined, whose empirical results suggest that grouping of diseases is very important to realize automated acquisition of medical knowledge from clinical databases. Thus, we focus on the role of coverage in focusing mechanisms and propose an algorithm for grouping of diseases by using this measure. The above example shows that rule induction with this grouping generates rules, which are similar to medical experts' rules and they suggest that our proposed method should capture medical experts' reasoning. This research is a preliminary study on a rule induction method with grouping and it will be a basis for a future work to compare the proposed method with other rule induction methods by using real-world datasets.

References

[1] Everitt, B. S., *Cluster Analysis*, 3rd Edition, John Wiley & Son, London, 1996.

[2] Pawlak, Z., *RoughSets*. Kluwer Academic Publishers, Dordrecht, 1991.

[3] Polkowski, L. and Skowron, A.: Rough mereology: a new paradigm for approximate reasoning. Intern. J. Approx. Reasoning **15**, 333–365, 1996.

[4] Quinlan, J.R., *C4.5 - Programs for Machine Learning*, Morgan Kaufmann, Palo Alto, 1993.

[5] *Readings in Machine Learning*, (Shavlik, J. W. and Dietterich, T.G., eds.) Morgan Kaufmann, Palo Alto, 1990.

[6] Skowron, A. and Grzymala-Busse, J. From rough set theory to evidence theory. In: Yager,R., Fedrizzi,M. and Kacprzyk, J.(eds.) *Advanes in the Dempster-Shafer Theory of Evidence*, pp.193-236, John Wiley & Sons, New York, 1994.

[7] Tsumoto, S., Automated Induction of Medical Expert System Rules from Clinical Databases based on Rough Set Theory. *Information Sciences* **112**, 67-84, 1998.

[8] Tsumoto, S. Extraction of Experts' Decision Rules from Clinical Databases using Rough Set Model *Intelligent Data Analysis*, 2(3), 1998.

[9] Zadeh, L.A., Toward a theory of fuzzy information granulation and its certainty in human reasoning and fuzzy logic. *Fuzzy Sets and Systems* **90**, 111-127, 1997.